



Estimating Heterogeneous Treatment Effect using Qausi-Experimental Designs: Evaluating the Impact of National Rural Health Mission (NRHM) on Maternal Health Outcomes in India

4927

Author: *Abu Afzal Tauheed*
Jawaharlal Nehru University
New Delhi-110067

Abstract

Estimating causal relation using observational survey data or non-experimental data has been a challenging task. Likewise, satisfying conditional independence over unobservable is an untestable and restrictive assumption. This paper uses machine and deep learning algorithms to resolve the fundamental problem of causal effect using non-experimental data. By satisfying a “sufficient” assumption of exogeneity the model identifies instances from the database that mimics the Quasi-Experimental Designs. This helps in estimating the complete distribution of counterfactuals for causal inference. The paper exploits the asymptotic properties – weak monotonicity and identity function – to show that average treatment effect (ATE) and conditional treatment effect (CATE) converges to true value and are consistent and unbiased. The assumption of connectedness of potential outcomes in two periods is central for model selection. The model uses the intelligence of system to look for units in population database and select that subset of data which satisfies the sufficient condition for casual inference (assumption of exogeneity). The study also suggests that an availability of well-defined database with extensive features and continuous time interval in developing countries will help to estimate the impact of key welfare schemes/programs. This paper addresses the casual effect inference for classification problem. The study estimates the heterogeneous treatment effect using differential eligibility rules of national health program (NRHM) over maternal health outcomes in India. It uses scenario of panel data from Indian Human Development Survey (2004-05 & 2011-12). The findings suggest that potency of intervention for increasing institutional delivery. However, the CATE for ante-natal (ANC), postnatal (PNC) and Safe-Delivery indicates that intervention failed to remove the inequalities.

JEL classification

C31; C33; C45; I18

Key Word: Heterogeneous Treatment Effect; Machine Learning; Causal Inference; Quasi-Experimental Design; Maternal Health;

DOI Number: 10.14704/nq.2022.20.10.NQ55471

NeuroQuantology 2022; 20(10): 4927-4959

1. Introduction

The literature on evaluation methods in three decades time has been voluminous, however, randomized controlled experiment remains the thumb rule for causal effect inference from a policy intervention (Athey and Imbens 2017). Randomize control experiment method assigns treatment randomly to individuals. It resolves the biasedness and missing data problem in estimation, and average treatment effect is a simple subtraction of average outcomes of non-treated group from treated ones. Intuitively, it means that the potential outcome of a



treated and non-treated individual would be same if the distribution of unobservable for both the individual is same. It also assumes that potential outcomes and treatment assignment are independent to each other. However, in practice conducting a randomized experiment becomes impossible due to financial, political and ethical consequences (Ibid.). For instance, it would be ethically wrong to abdicate any one from a surgery to study the impact of a surgery on people's health. As a consequence, large section of the empirical work on casual effect inference uses observational data¹, also known as non-experimental data². However, the challenge of using "observational data" for causal effect inference arises because of non-randomness of treatment assignment. Hence, potential outcomes and treatment assignment may not be independent³ in non-experimental data and may have common causes (Jensen, et al. 2008). The problem of missing data or fundamental problem of causal effect in observational data occurs, as treatment assignment may be dependent on potential outcomes. Intuitively this means that there are other co-founding factors (common causes), which influence the treatment assignment given the potential outcomes in post intervention period. Therefore, if one doesn't block the nodes of common causes that influence treatment assignment given potential outcomes it leads to a missing data problem⁴. The problem of missing data can be understood by simple example. Suppose, a database, observes 10 individuals for two time periods, before and after the policy intervention, among them 5 have received the treatment and rest are untreated. For estimating average treatment effect (ATE) one has to ensure that common causes which influence treatment and potential outcome has been blocked (Jensen, et al. 2008). In other words, the distribution of co-founding factors (covariates) is equal for treated and non-treated individuals group. This is known as selection of observables or strongly ignorable treatment assignment (Rosenbaum and Rubin 1984), meaning that all the variables related to potential outcomes and treatment assignments are controlled⁵ (Ibid). Therefore, to overcome the problem of missing data in non-experimental data⁶ (Rosenbaum and Rubin 1984) (Heckman, Ichimura and Todd 1998) (Imbens 2003) advocated reducing the dimension of observable features (co-founding factors). So, that the distribution of observable features for treated and non-treated individuals group are equally distributed⁷. An additional requirement of estimating ATE in observational data is to make sure that for every treated individual in the sample there is at least one non-treated individual whose distribution of observables are equal to that treated individual. This is known as overlap or matching assumption (Cameron and Trivedi 2005).

Notwithstanding, researchers often use set of assumption while using observational data known as "identification or empirical strategies" (Athey and Imbens 2006). For estimating

¹ Observational data are the survey data for households, individuals or for any units of interest. It is also considered as sample survey (Cameron & Trivedi, 2005).

² The study will use the term observational and non-experimental data inter-changeably.

³ In probabilistic model this assumption becomes crucial for estimating the conditional probability.

⁴ This is known as fundamental problem of causal effect inferences in evaluation literature.

⁵ This assumption is also known as un-confoundedness (Imbens 2003) (Athey and Imbens 2006).

⁶ This is also known as fundamental problem of causal effect inference.

⁷ This is known as model selection problem. The present study will extensively use this for model selection.

ATE the present study uses two set of assumptions; first set of assumptions assure that the missing data problem is resolved by blocking the common causes (nodes) between treatment assignment and potential outcomes (the study calls it a exogeneity assumption). In addition, the overlap assumption assures that for every treated individual there is at least one non-treated individual whose distribution of observable is equal to that of treated individual. Therefore the assumption of exogeneity and overlap consist the first set of assumption. In addition, the proposed study uses set of asymptotic properties, given the first set of assumptions, through which the model can estimate the complete distribution of counterfactuals⁸. Note, the present study uses the word counterfactual, which means what would be the outcome of treated individual if s/he has not been exposed to treatment, whereas what would be the outcomes of non-treated individual if they are exposed to the treatment. In other words, the present study proposes an optimization model selection method, which estimates the complete distribution of counterfactuals. Thus the proposed model can estimate the average treatment effect⁹, which is unbiased and consistent. The advantage of the present model is that it minimizes the empirical error and converges approximately to the generalized error¹⁰. Thus the study uses two sets of assumptions as an identification strategy for estimating the parameter of interest that is consistent, unbiased and robust¹¹.

The present study has used the discussed estimation method to calculate the average treatment effect (ATE) and heterogeneity in treatment effect (i.e. conditional average treatment effect) for National Rural Health Mission (hence forth NRHM) introduced in 2005 for improving the maternal health outcomes in respective states of India. In addition, this chapter will exploit the demand incentive cash benefit transfer¹² and regional based eligibility rule¹³ to identify the treatment effect of healthcare program- NRHM - on maternal health outcomes. The central focus of this chapter is to address the methodological aspect of policy evaluation using observational/non-experimental data. The discussed methodology will help to estimate heterogeneous treatment effect for average treatment effect (ATE) and conditional average treatment effect (CATE). It introduces to the *changes in changes* model

⁸ The central challenge for an evaluation method using observational data is to estimate the complete distribution of counterfactuals. Therefore, if can resolve the issue of missing data problem or fundamental problem of causal effect then the model can estimate the complete distribution of counterfactuals.

⁹ Estimating average treatment effect (ATE) has been the central challenge in observational data, as treatment is not assigned randomly. In addition, methods such as Difference-in-Differences (DID) becomes restrictive for estimating ATE, however, it can estimate average treatment effect on the treated (ATT) (Blundell and Dias 2009).

¹⁰ In other words, the model claims that the empirical error can be generalized to the population error. For example, if researcher wishes to estimate the impact of area size of houses over the price of houses, then the model shall be a good predictor for the houses not included in the sample.

¹¹ The proposed model is double robust as it combines two identification methods (matching and regression) for estimating the parameter of interest.

¹² The study utilize the information of conditional cash transfer received by women to distinctly identify them as treated and non-treated women. In other words, the study identify a woman to be treated if she has receive the treatment and to be non-treated if she has not received the cash benefit.

¹³ The study will use the information about regional based eligibility rule for the program to decompose the treatment effect. This is known as conditional average treatment effect.

for discrete outcome variables¹⁴ advocated by (Athey and Imbens 2006)¹⁵. Therefore, we first estimate the distribution function of average treatment effect (ATE) and then the conditional distribution of ATE at selective covariates to see how the average effect of the treatment varies between High Focus States (HFSs)¹⁶ and Non-High Focus States (Non-HFSs)¹⁷ and further by their respective regions identified as urban and rural areas. It finds that the program significantly raised the institutional delivery in rural regions of HFSs and Non-HFSs. However, the proliferation in “at least four ANC¹⁸ checkup” and PNC (Post Natal Care) had moderate ATE, however the conditional average treatment effect (CATE) for the rural areas in HFS has been lower relative too Non-HFS. The findings suggest maternal healthcare disparities between HFS and Non-HFS states and their respective regions still prevails.

The chapter is divided into five sections; second section explains the differential eligibility rule and potential outcomes of the program (NRHM) and, treated and comparison (non-treated/control) groups. Third section briefly explains the existing methodology and the proposed modified identification methodology. This section will introduce to the comprehensive evaluation model used for estimating parameter of interest. Likewise, the section will also explain the estimation method for heterogeneity in treatment effect. Fourth and fifth section explains the data and methodology and empirical results respectively. Last section concludes the chapter.

2. The Program: Eligibility Criteria for Treatment Assignment and Potential Outcomes

The National Rural Health Mission (NRHM) was a much-needed intervention in India for primary healthcare with special attention to Reproductive Child Health (RCH)¹⁹ care. The program had embodied key sub component to cater the healthcare utilization at primary healthcare system. Likewise, NRHM²⁰ enveloped Janani Suraksha Yojana (JSY)²¹ under RCH flexible pool for incentivizing the women in high focus states (henceforth HFSs) and Non-high focus states (henceforth Non-HFSs) with differential eligibility rule²². The program

¹⁴ This is also known as classification problem.

¹⁵ This model uses a set of assumptions for production function along with conditional independence assumption.

¹⁶ High Focused States (HFSs) includes Assam, Bihar, Chhattisgarh, Himachal Pradesh, Jharkhand, Madhya Pradesh, Odisha, Rajasthan, Uttarakhand, and Utter Pradesh. These states were given higher importance for achieving better health outcomes.

¹⁷ Non-High Focused (Non-HFSs) includes Andhra Pradesh, Gujarat, Haryana, Karnataka, Kerala, Maharashtra, Punjab, Tamil Nadu and West Bengal.

¹⁸ Ante Natal Care, the analysis considered at least 4 ANC check up as per the recommendation of WHO.

¹⁹ This is formally known as RCH II.

²⁰ The program has other sub-summed items to cater the primary healthcare with special reference to maternal and child health such as, Mission flexible pool (it includes NRHM flexible pool for hiring community health workers and Infrastructure maintenance fund for strengthening public health infrastructure); National Disease Control Program (for non-communicable diseases) and, IPPI (for polio eradication).

²¹ JSY was completely funded by central government to incentivize pregnant women to avail the formal healthcare services.

²² The program has modified the eligibility rule in the initial period (after first year of implementation), however, for the analysis the study has considered the final eligibility rule proposed by the government.

targeted to incentives all pregnant women in HFSs who wishes to avail deliver care in public health center or in credential private healthcare.

The eligibility norms for conditional cash benefit were simple: All pregnant women who deliver in public health center or in credential private health center where entitled to receive the cash benefits. The eligibility criteria would differentiate between HFSs and Non-HFSs states and, further amid rural and urban regions of the particular states. These regional disparities in eligibility were essentially brought for women's who resides in rural areas of laggard (HFS) states. For instance, women in rural part of HFSs were eligible to receive 1400 INR where as in urban areas it was 1000 INR. While on the contrary, women in rural region of Non-HFSs were eligible to receive only 700 INR where as in urban areas it was 600 INR²³ (Debnath 2012). An additional norm for eligibility was based on socio-economic factors. For instance in Non-HFSs, eligibility was limited to women from SC, ST and BPL category²⁴, where as in HFSs all women were eligible²⁵. Given the eligibility rule, the program can be identified as one of the "national implementation", being managed for everyone in the country who meets the eligibility criteria.

These eligibility norms in respective regions of the states determined the treatment group for the program. The program for maternal health was designed to provide safe motherhood²⁶ with an objective to reduce infant mortality and maternal mortality rate in the country. It is widely discussed in literature that adequate antenatal care is a prerequisite for reducing infant mortality and maternal mortality rate (McDonagh 1996) (Oyerinde 2013), likewise post-natal care are also essential for reducing IMR and MMR (Sines, et al. 2007) (Rai and Singh 2012). In addition immunization for the mother and newly born child is necessary for reducing IMR and MMR (UNICEF 2009) (Olusegun, Thomas and Micheal 2012). The program document identifies provision of antenatal care, institutional delivery, postnatal care, immunization and services related to mother and child healthcare as the expected outcomes at community level (GoMP 2006).

The program (NRHM) further motivated the community health workers (ANM, ASHA and ANW)²⁷ by offering them cash benefit for assisting the pregnant women for safe maternity. The community health workers (CHWs) was motivated by the payment of 600 INR in rural areas and 200 INR in urban areas of High Focus States, where as in Non-HFSs there was no provision of motivational payment (Debnath 2012)²⁸.

Notwithstanding, the Program (NRHM) had special sub-summed item entitled for infrastructure creation and maintenance fund. These funds were entitled for strengthening

²³ The government also incentivized community health workers (ANM, ASHA, ANW) for facilitating safe motherhood.

²⁴ SC, ST and BPL stands for schedule caste, schedule tribe and below poverty line respectively.

²⁵ In addition, the eligibility for women in Non-HFSs was restricted up to two live births, however the study has relaxed this norm for an analysis.

²⁶ By definition of safe motherhood includes four contributing factors such as antenatal care, delivery care, postnatal care and, immunization; it helps in reducing the risk of mortality in mother and child (WHO, Make Every Mother and Child Count, The World Health Report 2005).

²⁷ ANM, ASHA and ANW stand for auxiliary nurse midwifery, accredited social health activist and anganwadi workers respectively.

²⁸ A study conducted by Debnath (2012) asserted that additional payment to community health workers raised the probability of key maternal health outcomes.

the primary and secondary healthcare system. One of the major obstacle in utilization of maternal health services has been *sizeable geographical distance and inadequate healthcare systems; non-availability of community health workers, doctors, drugs and medicines, water, electricity, examination room, beds, OPD hours etc.* (Noordam, et al. 2011) (Panel 2010) (Singh 2016) (Kumar and Dansereau 2014) (Jacobs, et al. 2011). These supply side factors are necessary condition for the provision of adequate safe delivery²⁹ ultimately leading to a fall in IMR and MMR (Walraven, et al. 2000) (Jacobs, Judd and Bhutta 2016) (Frankenberg 1995). Moreover, large share of NRHM flexible pool funds used for recruiting additional community health workers (CHWs; ANM ASHA, ANW). These funds were also supposed to be utilized to offer them adequate training for assisting maternal and newly born child health services. Therefore, the program was launched all across the Indian states in 2005 with differentiated eligibility rules for treatment. Identification of the treatment effect in non-experimental data requires additional set assumptions for conducting an experiment. Given the program (NRHM) the study evaluates the impact of it at community level expected outcomes³⁰. The experimental evaluation analysis for this study is based on the data before and after the national roll out of the program.

2.1 Choice of the outcome variables

The study focuses on the impact of a program (NRHM) on proportion of eligible women completing four contributing elements – or components – of safe delivery also termed as safe motherhood³¹. The desire to investigate these components of safe delivery comes from the program identified community level expected outcomes defined in the policy document (GoI 2005) (GoMP 2006). The intervention aimed to reduce the infant mortality and maternal mortality rate by realizing the expected outcomes identified at the community level. It was recognized that the provision of adequate antenatal care (ANC), institutional delivery (ID), post-natal care (PNC) and immunization to eligible women were essential to meet the community level expected outcomes. Thus study primarily focuses on four-core component of safe motherhood i.e. ANC, ID, PNC, and Immunization.

However, for looking at the adequate/minimum level of expected outcomes for reducing maternal mortality rate, the study follows the definition stated by WHO. As per WHO recommendation a pregnant women should have at least four ANC checkup assessments³²

²⁹ The study have used the term “*Safe Delivery*” to indicate that a women who has received the entire contributing elements for safe motherhood i.e. Antenatal care, Institutional Delivery, Post Natal care and, Immunization. Thus the study considers a woman to have safe delivery if she has received all the contributing elements of safe motherhood. WHO in their training modules highlighted these stages to safe motherhood (WHO, Make Every Mother and Child Count, The World Health Report 2005).

³⁰ A program document from Government of Madhya Pradesh has asserted that an adequate ANC, Institutional Delivery, PNC and, Immunization is an elementary expected outcomes of this program at community level. These expected outcomes should be realized to eliminate the inter-state and inter-region disparities in maternal health outcomes, for ultimately reducing the IMR and MMR. (GoMP 2006) (GoI, 2005).

³¹ The study uses safe delivery and safe motherhood interchangeably.

³² The aim of providing at least 4 ANC is to reduce the risk of disease and pregnancy complication. The recommendation also highlights the provision of immunization i.e. Iron syrup, Folic acid syrup, Tetanus and, Vitamin A. The study will incorporate the aspect of immunization while looking at safe Delivery.

(WHO 2016). It is further advocated that a pregnant women should have a provision of PNC checkup within 24 hours of the delivery (WHO 2013). The study primarily looks at three core outcomes variables ID, ANC and PNC. For post natal checkup (PNC) the study will look at two aspects; first, in two months period time after the delivery did both - mother and child - has received any post natal care, and second, is the PNC was given within 24 hours after the delivery.

In addition, to the expected outcomes mentioned above, the study also estimates the impact of the program on safe delivery³³ i.e. what is the probability of treated group to have a safe delivery after the policy intervention relative to comparison/control groups. Incorporating the concept of safe motherhood – from WHO - the study constructed the outcome variable for safe delivery; it incorporates all the four contributing elements or components of safe motherhood i.e. ANC, ID, PNC and Immunization³⁴. Notwithstanding, the study proposes the hypothesis that the proportion of those women who has received all the four components of safe maternity in treated and comparison group.

In nutshell: conditional cash benefit given to a pregnant woman is considered to receive the treatment assistance through the intervention proposed by the NRHM program. The evaluation aims to measure the impact of maternal health intervention and improved healthcare system on the probability of having institutional delivery, at least four antenatal care, PNC within 24 hours of delivery and, postnatal care for both mother and newly born child within the period of two months after delivery among the groups. The study further estimates the changes in probability of having safe delivery among groups due to intervention. Furthermore, the study uses the differentiated eligibility rule between states, regions and economic class to estimate changes in conditional probability of the outcome variables among the groups. The analysis also predicts the probability differences within the treated group with differentiated eligibility rule.

3. Identification and Estimation Methods

The present section introduces to the existing identification methodology for estimating treatment effect. In particular, the study will explore the existing identification methods for observational or non-experimental data. The central problem for estimating treatment effect in observational data arises due to non-randomness of treatment assignment (Blundell and Dias 2009) and common causes between treatment and potential outcomes (Jensen, et al. 2008).

3.1. Existing Methodology

Let us suppose the following set up for estimating the average treatment effect with discrete treatment variable. For all unit of observation (individuals) in sample population there is two pair of potential outcomes $Y_i(w)$ where $W_i \in \{0,1\}$ is the treatment variable where 0

³³ The variable for safe delivery was constructed by using the definition of four contributing elements for safe motherhood.

³⁴ In other words, the variable is constructed in way, where a woman has been allotted the value of 1 if she has completed or received all four contributing elements of safe motherhood i.e. at least 4 ANC checkup, Institutional Delivery, PNC within 24 hours of birth, Immunization (that includes Iron and Folic syrup or tablets and Tetanus injection). If any of the mentioned component of safe motherhood is skipped in the maternity process then the women is allotted by the value of 0 (zero).

indicates the comparison group and 1 for treated group for $i = 1, \dots, N$ (Rosenbaum and Rubin 1984). So the model looks like,

$$Y_i = Y_i(w)$$

$$Y_i(w) = \begin{cases} Y_i(1) & \text{if } w = 1 \\ Y_i(0) & \text{if } w = 0 \end{cases}$$

average treatment effect (ATE)³⁵ and average treatment effect on treated (ATET) can be estimated as

$$\begin{aligned} \tau^{ATE} &= E[Y(1) - Y(0)] \\ \tau^{ATET} &= E[Y_i(1) - Y_i(0)|W_i] \end{aligned}$$

and conditional average treatment effect (CATE)

$$\tau(X) = \frac{1}{N} \sum_{i=1}^N E[Y(1) - Y(0)|X]$$

In general, applying standard DID settings over observational data produces several limitations. Firstly, the estimation based on standard DID settings may often results in over or under estimation of treatment effect³⁶ (Athey and Imbens 2006) (Blundell and Dias 2009). Likewise, with discrete outcome variable the standard DID model fails to estimate the treatment effect within bounded region and perhaps ranges beyond 0 and 1. Also, estimating a non-linear DID model with an assumption of linearity in error term (i.e. common trend and randomization) makes the estimation unreliable (Blundell and Dias 2009) (Athey and Imbens 2006). Similarly, using an index function with DID settings fails to account the heterogeneity in treatment effect. Moreover, the common trend assumption³⁷ and randomization of treatment assignment³⁸ is a prerequisite for estimating “average treatment effect” ATE. Intuitively, random assignment assumes that treated and non-treated individuals group is equal in all features apart from their treatment status (Blundell and Dias 2009). Likewise, both the assumption becomes strongly restrictive for estimate average treatment effect (ATE) in difference-in-difference (DID) settings (Blundell and Dias 2009). On the other hand, matching methods³⁹ intends to create set of treated group within the non-treated group. This helps is estimating the counterfactuals for solving the problem of missing data by assuming independence between potential outcomes and treatment. It uses weights score, based on the distribution of covariates that is associated with the outcomes (of treated and non-treated groups) and treatment assignments (Ibid, 44). The matching method under the assumption of un-confoundedness and overlap can estimate the ATE and ATET (Imbens 2003).

³⁵ Average treatment effect on treated can be estimated as $\tau = E[Y_i1 - Y_i0|W_i]$. One must keep in mind for estimating ATE randomness in treatment assignment is essential.

³⁶ A detail explanation of Difference-in-Differences model is beyond the scope of this study. However, interested readers may refer (Khandker, Koolwal and Samad 2010) (Blundell and Dias 2009).

³⁷ This assumption means that in absence of treatment both treated and non-treated individuals will experience similar trend in potential outcomes over the period of time.

³⁸ Randomization assumption is essential for estimating ATE. It indicates that for any policy intervention individuals should be chosen randomly for treatment assignment. Moreover, treatment assignment shall not be dependent of potential outcomes in time period one.

³⁹ For better understanding please refer to (Imbens 2003) (Blundell and Dias 2009)



Nevertheless, a mixed method, which combines the matching methods along with regression settings, may remove the biasedness and reduce variance (Imbens 2003). However, applying matching methods along with DID settings can only estimate ATET (Blundell and Dias 2009). The problems even become acute when one has to deal with classification problem in static settings⁴⁰.

3.2. The Changes in Changes Model

To overcome these challenges (Athey and Imbens 2006) (AI hence forth) introduces two groups and two-time period changes-in-changes (CIC) model that introduce the dynamic estimation method for mapping the average treatment effect⁴¹. The AI model nest the functional form⁴² assumption, which makes it distinct from standard DID settings. Similarly, the AI model assumes that the production function (the potential outcome function also known as hypothesis) $Y_i(w) = h_i(\cdot)$ is strictly monotonic⁴³ in unobservable. The monotonic assumption remains central to the estimation method proposed by AI as it assures the existence of inverse production function. Thus using the inverse transformation method⁴⁴ the model estimates the cumulative distribution function for different groups at different instances (time period). Therefore, the CIC model advocated by AI estimates the complete distribution of counterfactuals⁴⁵ for treated and non-treated individuals groups if and only if the assumption of conditional independence is satisfied. Therefore, the CIC model proposed by Athey and Imbens is a predictive model, which assumes the followings for discrete dependent variables:

- i. The production function $Y_i(w) = h_i(\cdot)$ is weakly monotonic in unobservable.
- ii. The production function $h_i(\cdot)$ is an identity function.
- iii. The distribution of unobservable varies across groups independent of time.
- iv. The distribution of unobservable is independent of groups given potential outcomes and time period i.e. $U \perp W|Y, T$; this is known as conditional independence assumption.
- v. The distribution of unobservable for treated group will be equal or subset to the distribution of unobservable in non-treated group i.e. $U_1 \subseteq U_0$.

Intuitively, the AI model maps the individual's group over the distribution of unobservable and it asserts that irrespective of treatment groups the outcome will be same for individuals if their distribution of unobservable is same. The first two assumptions make sure that the

⁴⁰ Classification problem arises when outcome variable can take limited discrete values.

⁴¹ The approximation method used by Athey and Imbens are reliable than the standard DID methods, for better clarification please refer to (Athey and Imbens, 2006).

⁴² The functional form assumption in standard DID settings assumes that in absence of treatment the outcomes of treated and non-treated will be same.

⁴³ The AI model assumes strictly monotonic production for continuous outcome variable, however for discrete dependent variable they assume weak monotonic production function.

⁴⁴ Inverse Transformation Method can be used when we have an identity function i.e. the function behaves monotonically and invertible.

⁴⁵ Counterfactuals for treated individual's group indicates the distribution of potential outcomes if they were not given the treated, similarly, the counterfactuals for non-treated groups indicates the distribution of outcomes if they were exposed to treatment.

production function is well defined and ranges within the bounded region. It assures that the limit probability and cumulative distribution function converges and the estimated parameter is consistent. Similarly, fourth and fifth assumption resolves the fundamental problem of causal effect or missing data. Intuitively, it is similar to the assumption of unconfoundedness and overlap assumptions discussed above.

The proposed **AI** model approximates better average treatment effect than standard DID settings however, the model maps the individuals by estimating the distribution of unobservable⁴⁶. Intuitively, it means that the model map the potential outcomes from unobservable information of individual's group as time changes. The model maximizes the production function based on unobservable elements of individuals in the sample set⁴⁷. Therefore the proposed model by **AI** works better (compare to standard DID), as it approximate the average treatment effect based on unobservable information in sample data set. Intuitively, this indicates that **AI** model works better as it works for large data set, mapping the individual based on their unobservable information as time changes. It can account for heterogeneity in treatment effect as it estimates the treatment effect for every individual in the sample. However, the model may over estimate the approximation of treatment effect as it work on unobservable information in the data. In addition, satisfying conditional independence in unobservable is untestable, complicated to validate and seems far than sufficient for estimating heterogeneous treatment effect.

3.3. Recent Development in Causal Inference using Machine Learning

A recent development in causal effect inference, which combines the predictive methods with fundamental law of causal inference for resolving the problem of missing data (i.e., accounting the complete distribution of counterfactuals) use the estimation tool from machine learning (Athey and Imbens 2017). There are two kinds of learning method in machine learning tools i.) Unsupervised and ii.) Supervised learning. A recent development in casual effect inferences uses supervised learning to solve the missing data problem. The advantages of using supervised learning to learn the counterfactual distribution is that one don't have to verify the assumptions because the assumptions are based on sample data which give us the training sample (Pearl 2009) (Ng, Supervised Learning 2008) (Ng, Learning Theory 2008) (Ng, Regularization and model selection 2008). The primary objective for any model selection methods remains to remove biasedness (i.e., estimating the complete distribution of counterfactuals) and to reduce the variance in estimation. In other words, the elementary reasoning to use supervised learning is to minimize the empirical error that represents the generalized error (Ng, Learning Theory 2008). Let us suppose we have a sample set which consist the information about 100 individuals, where Y_i is a random discrete outcome variable that takes the value 1 if the person has cancer and 0 otherwise; and X_i is a random input variable that takes the value 1 if s/he is a smoker and 0

⁴⁶ The central argument of model is that the outcomes of individuals are mapped through groups, time and unobservable characteristics. In other words one can understand this by a simple example; suppose we have a linear function $Y=f(x)$ if it satisfies the monotonic condition and is invertible then the inverse of this function will give us the distribution of unobservable in Y , which is unexplained by x i.e. $u=f^{-1}y$, where u indicates the unobservable elements.

⁴⁷ This is mainly because the AI model assumes that the production function increases monotonically in U i.e. unobservable.

otherwise⁴⁸. Now the objective is to estimate the probability of cancer (effect) given the information does the person smokes or not (cause).

For estimating the causal relation between cancer and smoking, the major challenge remains to block all the nodes of common causes (between smoking and cancer) and mimic quasi-experimental design (QED) (i.e. randomness) to minimize the empirical error that converges to generalized error. Supervised machine learning helps in estimating causal effect between smoking and cancer, by identifying the units of observation in database that blocks all the nodes of common causes⁴⁹ between smoking and cancer (representing QED) which removes the biasedness and variation in estimating treatment effect. Initial task is to select a model that represents QED minimizing the empirical error and second, that empirical error shall mimic the generalized error (i.e. it should have internal and external validity).

There are two ways to minimize the distance between empirical error and generalized error. First is cross validation⁵⁰ that divides the sample randomly into k training sample⁵¹ and estimate the parameter of interest for k training sample (Ng, Regularization and model selection 2008). It estimates the causal effect based on the averages of k training sample. The other way to do this is through feature selection. Now suppose, from the above example, that there are J numbers of features⁵² available in the sample for smokers and non-smokers. The challenge for this model is to incorporate the J features in model that represent QED and reduce the empirical error. Due to dimensional issues the estimation method may over estimate the prediction and the distance between empirical and generalized error may go wider. Nevertheless, a feature selection method may reduce the dimension of features and identifies the features that are sufficient to mimic QED, which can be retained for the analysis. For the sample of 100 units of observation suppose we have 10 features for each unit in database. To solve the model selection problem one can think as follows; suppose from the given 10 features, we first reduce the dimension by using any standard score method (for example, propensity score method) for 4 features, then for 7 and further for 9 features iteratively (Jensen, et al. 2008). Using this filter feature selection method one can select the automated QED (ibid, 375) that solves the model selection problem. From these three available reduces features, the model selects that reduced features through iteration that has the highest score (propensity score) and minimizes the biasedness. Further it reduces the distance between empirical and generalized error using classification machine learning algorithms such as decision tree, random forest or support

⁴⁸ One can also include Xi2 to estimate probability of cancer as the number of smoking cigarettes increase.

⁴⁹ Common causes could be diet, physical activity, exposure to radiation, viruses and infections, genetics and others.

⁵⁰ Formally, this is known as k-fold cross validation.

⁵¹ In Machine Learning the term training sample indicates the selected unit of observation to estimate the parameter of interest from database.

⁵² One may think of these features such as age of the smoker, number cigarettes smoked by the person etc.

vector machine. Importantly, applying cross validation after feature selection and classification algorithm improves the approximation method.

3.4. Identification of Changes-in-Changes Model using Machine Learning

Comparing the two ways (cross validation and feature selection) to overcome the model selection problem suggests that both has its own pros and cons. However, the present study advocates for feature selection method for solving the model selection problem as it maximises on observables and minimizes the empirical error for reducing its distance from generalized error. Therefore, going back to the example with 100 individual where the objective is to find causal effect between smoking and cancer. Using iterative feature selection method, the model may selects 70 unit of observation (out of 100 unit of observation) on the basis of 7 reduced features or characteristics (with highest propensity score in all 10 features). By doing this the supervised learning find those instance in the database that blocks all the nodes/stream of common causes between treatment assignment and potential outcomes (Jensen, et al. 2008) (Pearl 2009) and mimic the quasi-experimental design on which one can infer the cause-effect relationship and can claim internal and external validity (Jensen, et al. 2008). In other words, the model selects these 70 individuals, where there is no confoundedness between cause and effect. Hence the training sample represent that sub-sample data set from database that identifies the instances where treatment assignment is randomized (representing QED). Intuitively, it indicates that now for every smokers the training sample consist at least one non-smoker who has similar distribution of 7 features. Coming back to causal effect inference, the feature selection method selects the independent training sample where treatment assignment is randomized by blocking all the nodes/streams that may influence the treatment assignment and potential outcomes. Therefore, using the machine-learning tool one can resolve the model selection problem and chooses that training sample that satisfies the sufficient assumptions of un-confoundedness and overlap. Accordingly, by using machine learning one can estimate the complete distribution of counterfactuals for both treated and non-treated.

Consequently, using a production function, which satisfies the asymptotic properties one can estimate the average treatment effect and conditional average treatment effect. Therefore, optimizing the production function which is weakly monotonic (as we are dealing with classification problem) in X_i the model minimizes the training error to approximate it to the generalized error.

The Model

Let $Y_{i,(w,t)}$ be the potential outcomes for individuals $i = 1, \dots, N$; for group $W \in \{0,1\}$ over time period $T \in \{0,1\}$. Let us also assume that $Y_{i,(w=0)}$ be the counterfactual for those individuals who did received the treatment, whereas $Y_{i,(w=1)}$ indicates the counterfactuals for those individuals who didn't receive the treatment. This model has a special characteristic for predicting changes in changes after an intervention between treated and non-treated groups. Let us assume that the model wishes to predict the success of potential outcomes given the following production function:

$$Y_i(w, t) = h_i(W, T, X; M, \tau, x)$$



where, $h(\cdot)$ represents the production function, where as W and T indicates treatment group and time period; M is an interaction vector between groups and time periods; X is a random vector of observable features and, τ (tau) is the parameter of interest for predicting changes in potential outcomes given the change due to intervention.

The present approach advocates for casual effect identification for classification problem i.e. potential outcome may only take discrete values. To be specific, the study makes a case when dependent variables take binary value 0 and 1.

The proposed approach attempts to link the potential outcomes to groups of individuals, time, and observational elements from a non-experimental database. In addition, this model estimates the distribution of unobservable characteristics that maximizes the likelihood of outcomes⁵³.

One of the main reasons for biasedness in evaluation process is due to non-stochastic sample data. The model creates a condition that makes the sample data to behave stochastically. One of the ways to do this is to reshape the data, which satisfies the stochastic condition and gives a training sample. In other words, from the available database, the model selection method picks those units of observation that satisfies the sufficient condition for causal effect i.e. it blocks all the nodes or common causes between treatment assignment and potential outcomes (Jensen, et al. 2008) making the training data set quasi-random or representing a Qausi-experimental design QED.

For making it applicable the model considers a function, which is called “Sigmoid function” – S function – that not only controls the potential outcome with respect to observable characteristics but also maintain a balance between time, based on different individuals groups at different instances. So if one assumes a random variable like $y_i = f_i(v, q)$, where v and q are mutually dependent which has a unique solution. Thus the probabilities $P_r(q)$ and $P_r(v)$ adhere with its identity to lie from zero to one. Furthermore, the model creates a scope to account the heterogeneity in treatment effect.

Assumptions

1. Exogeneity: This assumption states that all the common cause have been blocked which makes the treatment assignment dependent on potential outcomes i.e. treatment assignment are quasi-randomly allotted and are independent of potential outcome. By satisfying, this assumption one may justify the sufficient condition for randomness in training sample set.

The assumption of exogeneity is based on the following assumptions.

- i. Ignorability Assumption:

$$Y_0 \perp W | X_i$$

(where Y_0 is outcome at $T = 0$) this mean that outcome in pre-intervention period is independent of treatment assignment condition over set of observable features. It is implicitly assumed in this assumption that Y_0 and Y_1 are connected⁵⁴.

⁵³ This means that the model minimizes the unobservable (training error term) by maximizing the objective function given the observable information.

⁵⁴ Assuming the connectedness between Y_0 and Y_1 the model assume that any unobservable confounder U_y that causes Y_1 also causes Y_0 . Let us suppose the following causal relation Chain: $W \rightarrow X \rightarrow Y$ or $X \rightarrow W \rightarrow Y$ and Fork: $W \leftarrow X \rightarrow Y$ in any of these relation we can X -d separates W and Y by



ii. Overlap:

$$0 \leq p(x) \leq 1$$

by this assumption model assures that for every treated individual there is at least one non-treated individual with same observable features. For resolving the issue of dimension to map individuals from both the groups the model reduces the dimension by calculating $p(x)$ from X_i number of features.

2. The production function satisfies the properties of *sigmoid function*, which converges in limit probability and distribution. In other words the function satisfy the weak monotonicity assumption i.e. it is non-negative and non-decreasing function on observable information.
3. The production function is an identity function. An identity function is a function, which is invertible. In others words the function can be define in inverse. By definition a function is invertible if it is an identity function and satisfy the property of bijection. Suppose we have an identity function i.e.

$$Y = F_x(X)$$

it is invertible in X, then

$$X = F_y^{-1}(Y)$$

as we know that $Y = F_x(X)$,

$$X = F_x^{-1}(F(X))$$

Therefore by using identity function one can simply prove that it is invertible.

4. It is assumed that the training sample of treated groups will be equal or sub-set of non-treated groups for both instances.

$$S_1 \subseteq S_0$$

Furthermore, the model assumes that the distribution of unobservable will be independent of time given groups i.e. $U \perp T|W$. This assumption gets satisfied if one has strongly balance panel data.

Let us explore how sigmoid function satisfies the asymptotic properties. The following proposition will support the asymptotic properties for sigmoid function or *S function*.

Proposition 1: Let us suppose that the proposed production function follows the characteristics of sigmoid functions (*S function*).

Suppose,

$$h(x) = \frac{1}{1 + e^{-x}}$$

let us claim that this function is stochastically bounded and any estimation based on this function will surely be bounded in local maxima and local minimum.

$$h(x) = \frac{e^x}{e^x + 1}$$

blocking the directed path from W to Y, and they are statistically independent given X i.e. $(Y \perp W|X)$. Intuitively, the model assume that any sub-component U_{y1} of U_y that causes Y_1 also causes Y_0 and causal effect of such Y_0 cannot be zero. Given this the only possible way is to find the path which blocks the nodes/stream between W and U_y , alternatively if U_y is constant then W and Y_1 are unconfounded. In both case there is no confounding paths. So identifying a subset of database in which W is independent of Y_0 is either because U_y is constant or it has no effect on W, representing QED.



let

$$e^x = q$$

so,

$$h(x) = \frac{q}{q + 1}$$

let us take an counter example that $q = 5$ so,

$$\begin{aligned} h(x) &= \frac{5}{5 + 1} \\ &= \frac{5}{6} < 1 \cong 0.5 \end{aligned}$$

now let $q = -5$

$$h(x) = \frac{-5}{-4} \cong 1.25$$

since our interest lies in defining bounded-ness between 0 and 1. So limiting these upper and lower bounded-ness between 0 to 1, we have

$$0 \leq h(x) \leq 1$$

So,

$$h(x) : f(e^x) \rightarrow [0,1]$$

there are higher probabilities that the real value of the function always lies between zero to one. Thus, the probability of any function that belong to sigmoid family is surely bounded between zero to one. So the probability distribution function (PDF, hereafter) will be stochastically bounded from 0 to 1, which is non-negative and non-decreasing in nature (i.e. weakly monotonic) ■

Furthermore, if one put a constraint on sigmoid function $\{h(\cdot)\}$ to estimate probability mass function – for discrete outcome variable – then there is no limit defined on upper bound i.e. the probability cannot exceed from one. So we put the constraint on lower limit so that *S function* has tight bounds.

Additional features of *S function*:

Let,

$$h(x) = \frac{1}{1 + e^{-x}}$$

Differentiating both the side with respect to x , we have;

$$\begin{aligned} h'(x) &= \frac{d}{dx} \left(\frac{1}{1 + e^{-x}} \right) \\ &= \frac{1}{(1 + e^{-x})^2} (e^{-x}) \\ &= \frac{1}{1 + e^{-x}} \left(\frac{e^{-x}}{1 + e^{-x}} \right) \end{aligned}$$

this can be simplified as

$$= \frac{1}{1 + e^{-x}} \left(1 - \frac{1}{1 + e^{-x}} \right)$$

as defined,

$$h(x) = \frac{1}{1 + e^{-x}}$$

⇒

$$h'(x) = h(x)(1 - h(x)) \quad (*)$$

Equation (*) will be used in further derivation.



Proposition 2: If S function satisfies the weak monotonic condition. Then it can be shown that it is invertible function.

Suppose we have the following functional form

$$h(x) = \frac{1}{1 + e^{-x}}, \quad \text{where } x > 0$$

let, $y = \frac{1}{1 + e^{-x}}$ and $y = f(x)$

Solving this for x in terms of $y \in (0,1)$ implies

$$\begin{aligned} 1 + e^{-x} &= \frac{1}{y} \\ \frac{1}{e^x} &= \frac{1}{y} - 1 \\ e^x &= \left(\frac{y}{1-y}\right) \\ x &= \text{Log} \left(\frac{y}{1-y}\right) \end{aligned}$$

this implies,

$$\begin{aligned} f^{-1}(y) &= \text{Log} \left(\frac{y}{1-y}\right) \\ &= \text{Log} \left(\frac{Pr}{1-Pr}\right) \end{aligned} \quad \text{(A)}$$

where Pr indicates the probability. Equation (A), $\text{Log} \left(\frac{Pr}{1-Pr}\right)$, is well defined for our assumed range (0,1). It indicates the function $h(x)$ is invertible using *Inverse Transformation Method* (ITM). Thus S function $[h(x)]$ is invertible ■

Corollary: From equation "A" i.e. $f^{-1}(y) = \text{Log} \left(\frac{y}{1-y}\right)$ one can see that the probabilities are for the discrete values of y . Moreover, equation "A" satisfies all the properties of cumulative distribution function (CDF). In other words, it is well defined in range of [0,1].

3.5. Identification of Cumulative Distribution Function (CDF) over observable and unobservable features.

In this sub-section the study will explore how to estimate CDF for respective groups and time period that are strictly bounded and the production function is invertible.

Let,

$$\begin{aligned} Y_{i,(w,t)} &= h_i(W, T, X; M, \tau, x) \\ &= \frac{1}{1 + e^{-(\beta W + \lambda T + \tau M + \delta X)}} \end{aligned}$$

Also let $(\beta W + \lambda T + \tau M + \delta X) = f_i(g(\tau))$, where $f(\tau) = (W, T, M)$ and $g(\tau) = (W, M, X)$ are the dimensions. If X is affecting the treatment effect along with other observable factors then the model shows that $f(g(\tau))$. Thus τ is an arbitrary variable. In addition, the interest of the model is to estimate the consistent parameter τ that helps in estimating the average treatment effect (ATE) and conditional average treatment effect (CATE).

So,

$$Y_{i,(w,t)} = \frac{1}{1 + e^{-f_i(g(\tau))}}$$

for simplicity omitting suffix i .

$$Y_{(w,t)} = \frac{1}{1 + e^{-f(g(\tau))}}$$

by using equation "A" it can be written as



$$\begin{aligned}
 f(g(\tau)) &= \log \log \left(\frac{Y_{(w,t)}}{1 - Y_{(w,t)}} \right) \\
 &= \log \log \left(\frac{Pr}{1 - Pr} \right) \tag{B}
 \end{aligned}$$

Therefore, from *Proposition 2* we can show that our production function is invertible using *ITM*. Moreover from *Proposition 1* one can show that $h(\cdot)$ satisfies weak monotonic condition.

Let us drive a cumulative distribution function (CDF) for $W = w$ and $T = t$, which satisfy weak monotonic properties and it is invertible. So redefining the probabilities for treated and non-treated individuals, which are qausi-random. Most importantly, the model can estimate the unbiased and consistent probabilities if and only if the samples mimic quasi-experimental design and input variables are stochastic.

Identification of CDF with covariates (one can easily generalize this derivation without covariates) for $W = w$ and $T = t$, where P is a symbolic notation of probability (P_r).

$$\begin{aligned}
 P(W, X; M, \tau, x) &= h_{i,(w,t)}(W, T, X; M, \tau, x) \\
 P(W, X; M, \tau, x) &= 1 - h_{i,(w,t)}(W, T, X; M, \tau, x)
 \end{aligned}$$

where w and w' indicates the treated and non-treated groups respectively. The above equation could be written in the following compact way.

$$p(W, X; M, \tau, x) = \{h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{Y_{i,(w,t)}} * \{1 - h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{1 - Y_{i,(w,t)}}$$

So the CDF for training sample would be

$$\begin{aligned}
 P(W, X; M, \tau, x) &= \prod_{i=1}^n \{h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{Y_{i,(w,t)}} \\
 &\quad * \{1 - h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{1 - Y_{i,(w,t)}}
 \end{aligned}$$

As the information about the individuals group are independently (stochastically) drawn so the likelihood function would be

$$\begin{aligned}
 L(\tau) &= \prod_{i=1}^n p(W, X; M, \tau, x) \\
 &= \prod_{i=1}^n \{h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{Y_{i,(w,t)}} * \{1 - h_{i,(w,t)}(W, T, X; M, \tau, x)\}^{1 - Y_{i,(w,t)}}
 \end{aligned}$$

Taking log both the side

$$\begin{aligned}
 l(\tau) &= \text{Log}(L(\tau)) \\
 l(\tau_i) &= \sum_{i=1}^n \left[\{Y_{i,(w,t)} \log \log h_{i,(w,t)}(W, T, X; M, \tau, x)\} \right. \\
 &\quad \left. + \{(1 - Y_{i,(w,t)}) \log \log (1 - h_{i,(w,t)}(W, T, X; M, \tau, x))\} \right]
 \end{aligned}$$

Now maximizing the log likelihood function. Since, it is a randomized distribution one can differentiate the log likelihood function for finding the changes-in-changes (CIC). Here the model considers that the policy intervention came up for some specific time period. So for finding – CIC – the probability of an outcome it simply multiply the x in the result after omitting x at this place.

So,



$$\frac{\partial(\tau_i)}{\partial\tau_i} = \sum_{i=1}^n \left[\left\{ \frac{Y_{i,(w,t)}}{h_{i,(w,t)}(W, T, X; M, \tau)|X} - \frac{(1 - Y_{i,(w,t)})}{(1 - h_{i,(w',t)}(W, T, X; M, \tau))|X} \right\} * \left\{ \frac{\partial(h_{i,(w_j,t)}(W, T, X; M, \tau))}{\partial\tau_i} \right\} \right]$$

where w_j is the common individual from w to w' .

$$= \sum_{i=1}^n \left[\left\{ \frac{Y_{i,(w,t)} \{ (1 - h_{i,(w',t)}(W, T; M, \tau)) | X \} - \{ (1 - Y_{i,(w,t)}) (h_{i,(w,t)}(W, T; M, \tau) | X) \}}{\{ (h_{i,(w,t)}(W, T; M, \tau) | X) \} \{ (1 - h_{i,(w',t)}(W, T; M, \tau)) | X \}} \right\} * \left\{ (h_{i,(w,t)}(W, T; M, \tau)) * (1 - h_{i,(w',t)}(W, T; M, \tau)) | X \} * \left\{ \frac{\partial(\beta W + \lambda T + \tau M + \delta X)}{\partial\tau_i} \right\} \right]$$

Focusing our interest only on common characteristics being derived from both groups.

$$= \sum_{i=1}^n \left[\{ Y_{i,(w,t)} (1 - h_{i,(w_k,t)}(W, T; M, \tau) | X) - \{ (1 - Y_{i,(w,t)}) h_{i,(w_k,t)}(W, T; M, \tau) | X \} \} M_i \right]$$

On some given X , we have

$$= \sum_{i=1}^n \{ Y_{i,(w,t)} - h_{i,(w,t)}(W, T; M, \tau) \} M_i \tag{A'}$$

equation (A') can be understood as an output through which we can learn how observable and unobservable are distributed among individuals for different groups. It could be called as *learning equation*. The model optimizes the information provided by observable characteristics of an individual and simultaneously minimizes the unobservable (error term). Interestingly, equation (A') justifies that there exist estimation for parameter τ that maximizes the log likelihood function. In addition, it chooses that parameter $\hat{\tau}$ at which there exist a local maxima, i.e. $\frac{\partial(\tau)}{\partial\tau_i} = 0$. This indicates that the estimate parameter is consistent. Equation (A') helps to identify the cumulative distributed function (CDF) of $Y_{i,(w,t)}$, given the interaction between groups on time. In other words, it can identify the distribution of $Y_{i,(w,t)}$ i.e. $D_{y,wt}$ given the information of M, X and τ .

Note: The derivation for reaching at equation (A') will not loose its essence if one omits the observable characteristics X . However, accounting the observable characteristics will not only helps to maximize the probability of potential outcome but also minimizes the unobservable (training error). In addition the derivation could easily be used for multiple groups and time period.

Statement: Decrypting the fact about how equation (A') is uni-formally distributed across individuals group.

Let us assume⁵⁵

$$y = D(x) \text{ and}$$

⁵⁵ Here D represents the respective distribution



$y = h(x)$ is given by definition, where $h(x)$ is invertible.

So,

$$x = h^{-1}(y)$$

Now putting this in $y = D(x)$ gives,

$$y = D(h^{-1}(\hat{y})) \quad (C)$$

This gives the distribution of unobservables in $Y_{i,(w,t)}$, where treatment is not confounded by unobservable. In other words, equation (A') can estimate the distribution across individuals group. Furthermore, equation (A') helps in accounting the heterogeneity in treatment effect across individuals over the period of time.

Now generalizing equation "C" for any group of individuals and any instance of time. So,

$$D(h^{-1}(\hat{y})) = D_{u,w}(h^{-1}(y; t))$$

Thus, $D_{u,w}(h^{-1}(y; t))$ is the cumulative distribution function (CDF) for unobservable across individuals group. Now focusing on all four combinations of different groups and time instances.

a. For group $W = 0$ and $T = 0$

Given,

$$\begin{aligned} y &= h(W, T, X; M, \tau, x) \\ \Rightarrow y &= h(0, 0, X; \tau) \end{aligned}$$

Using the statement we can write the above equation as follows;

$$h(0, 0, X; \tau) = D_{0,X;\tau}(h^{-1}(h(0, 0, X; \tau)))$$

Applying CDF on both the side gives us the following

$$D_{y,00}(h(0, 0, X; \tau)) = D_{0,X;\tau}(D_{0,X;\tau}(h^{-1}(h(0, 0, X; \tau))))$$

as it was shown in proposition that the S function is invertible. This implies

$$D_{y,00}(h(0, 0, X; \tau)) = D_{0,X;\tau}(0, 0, X; \tau)$$

Now applying inverse transformation method (ITM) on CDF i.e. $D_{y,00}^{-1}$

$$D_{y,00}^{-1}(D_{y,00}(h(0, 0, X; \tau))) = D_{y,00}^{-1}(D_{0,X;\tau}(0, 0, X; \tau))$$

it implies,

$$h(0, 0, X; \tau) = D_{y,00}^{-1}(D_{0,X;\tau}(0, 0, X; \tau)) \quad (J)$$

b. Now applying the same fact for group $W = 0$ and $T = 1$ gives us the following;

$$D_{(0,0,X;\tau)}^{-1}(D_{y,01}(y)) = h^{-1}(y; 0, 0, X, \tau) \quad \forall y \in \{0,1\} \quad (K)$$

Now combining equation "J" and "K"

$$h(h^{-1}(y; 0, 0, X, \tau)) = D_{y,00}^{-1}(D_{y,01}(y)) \quad (L)$$

c. Applying the same fact for group $W = 1$ and $T = 0$ gives us;

$$D_{y,10}(h(0, 0, X; \tau)) = D_{00,X,\tau;1}(h(0, 0, X; \tau)) \quad (M)$$

Combining equation "L" and "M" further, applying the same fact for different group at different time.

$$D_{y,11}(y) = D_{u,1}(h^{-1}(y; 0, 0, X, \tau; 1))$$



$$\begin{aligned}
 &= D_{y,10}(h(h^{-1}(y, 0, 0, X, 1)0)) \\
 &= D_{y,10}(D_{y,00}^{-1}(D_{y,01}(y)))
 \end{aligned}$$

After this transformation our next result predicts the changes in potential outcomes given the available observable data, which thereafter finds, the flaws of the model – based on analyzing the result. This suggests that, there is always a possibility to optimize the estimation so that the training error mimics generalized heuristic.

Deduction: Inverse Transformation Method trigger us to deduce the average treatment effect (ATE):

$$\begin{aligned}
 ATE &= E[Y_{11}^M - Y_{11}^{A-M}] \\
 &= E[Y_{11}^M] - E[\alpha(Y_{10})] \\
 &= E[Y_{11}^M] - E[D_{y,01}^{-1}(D_{y,00}(Y_{10}))]
 \end{aligned}$$

where α is an approximation rate which converges in distribution. In addition conditional average treatment effect (CATE) can be identified as follows.

$$CATE = E[(Y_{11}^M - Y_{11}^{A-M}) | X = x^*]$$

Formally, this could be written as,

$$CATE = \frac{1}{N} \sum_{i=1}^n E[(X_i = x^*)] - E[(Y_{i,(11)}^{A-M} | X_i = x^*)]$$

Thus the proposed model looks for the information from sample and select those units for estimating the parameters that optimizes on observables – simultaneously minimizes the error term- and choosing that probability that maximizes the occurrence of potential outcomes. Therefore the derivation, suggests that by guiding the system we can exploit the “given intelligence of the system” to identify those unit for analysis that maximizes the occurrence of outcome by optimizing the available observable information for the unit of measurement. The calculation advocates that the estimated parameter $\hat{\tau}$ is *consistent* if it follows a set of conditions (See. Assumptions) and it’s unbiased as it solve the problem of missing data for calculating the complete distribution of counterfactuals.

3.7. Data and Methodology

For empirically analyzing the model, the study uses the observational survey data provided by Indian Human Development Survey (IHDS)⁵⁶. IHDS (observational survey) data provides wide range of information for both household and individuals over the period of time. The study combines both set of information from the survey to make the sample set informative. In addition, IHDS tracks down the household and individuals to see how political, economic

⁵⁶ Desai, Sonalde, Reeve Vanneman and National Council of Applied Economic Research, New Delhi. Indian Human Development Survey (IHDS), 2005. ICPSR22626-v8. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-06-29. <http://doi.org/10.3886/ICPSR22626.v8>.
 Desai, Sonalde, Reeve Vanneman and National Council of Applied Economic Research, New Delhi. Indian human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-07-31. <http://doi.org/10.3886/ICPSR36151-v2>



and social factors – both at macro and micro level- affects the households and individuals over the period of time. Moreover, IHDS is the only sample survey in the country which provides the panel survey with attrition rate close to 15 percent. For analysis, the study uses two rounds of IHDS data – IHDS-I (2004-05) and IHDS-II (2011-12) - for accounting the changes in potential outcomes (see section two) over the period of time due to change in policy intervention.

The study tracks all the eligible women – with reasonable observable features of individuals and households - for both the rounds. It was found that there were 4456 eligible pregnant women whose information has been collected in both the rounds (for panel data the study has 8912 unit of observations). As the model suggest if one does an analysis over this sample then the estimation may turn out be inconsistent and biased. Leaving little scope for us to rely on the estimation. Interestingly, equation (A) suggests the system to learn for estimands, which maximize the probability, once the assumption of exogeneity is satisfied through feature selection. To understand this consider an example of twin problem:

3.8. Twin Problem for Model Selection

Let us suppose a pharmaceutical company wishes to introduce a drug to diagnose cancer patients. The central challenge for the company is to estimate the effective treatment effect due to drug usage. In general, although incorrectly, it is argued that if we compare the outcomes for two set of individuals – one who have been diagnosed by that drugs and the other set of people who are not exposed to drug – by simply taking the average differences in the outcome may give us undesired and biased results (Hernán and Robins 2017). The company faces two kind of problem; *first* how to estimate the effective treatment effect and *second*, it wishes to choose that model from equation (A), which maximizes the probability for effective treatment effect i.e. minimizing the distance between empirical and generalized error. In other words, the system runs "v" number of models through equation (A) to learn the consistent and unbiased estimands that converges to true (natural) unknown parameter τ . The objective is to learn that particular model from equation (A), which minimizes the cost (training error). There are two ways to do it first is cross validation or *k-fold* cross validation, while second deals with feature selection. Through feature selection, the system heuristically searches for units of observation in database that identifies instances which represents qausi-experimental designs (QED)⁵⁷. Let us suppose the sample set has $k = 1, \dots, k$ number of observable features in database. However, the important task remains to identify the key features that are relevant to learn the optimal model from equation (A). Now if one has k number of features, then the selection problem for the model becomes acute, as there would be 2^k possible model to select from. This is known as selection problem (Rosenbaum and Rubin 1984) (Hernán and Robins 2017). As a consequence one can reduce the dimension for heuristic search for the optimal model selection. Therefore, the system learns and chooses a model that represents better approximation for average treatment effect. In other words, it is argued, that if S_i denoted

⁵⁷ This blocks all the confounding factors (both observable and unobservable) between treatment and potential outcomes.



as sample set where the objective is to choose that sub-sample from $V_i = v_1, \dots, v_d$ which represents QED (having both internal and external validity (Jensen, et al. 2008)) thereby reducing the distance between training and generalized error representing better approximation method. As the study showed in previous section that satisfying the exogeneity assumption is central for casual inference. Therefore, if one does the feature selection (by reducing the dimension of k features) based on past information- about health and other observable characteristics for treated and non-treated individuals (assuming that the potential outcomes in both time period is connected), then the following identification method can be used to estimate effective treatment effect.

The present study intends to follow the same procedure for calculating the average treatment effect (ATE) and conditional average treatment effect (CATE). The study uses strongly balanced panel data for 4456 units of individuals – before and after the intervention (NRHM) – for two-time period. Furthermore, for optimal model selection the study reduced the dimension of key features for individuals to compare the treatment effect between the treated and non-treated. The system reiterate and selects seven key features of a woman for reducing the dimension such as – Age, Level of Education, Sector (URBAN/RURAL), Caste, Religion, Assets Holding, Economic class (BPL/APL) – for comparing the treated individuals with non-treated one's. Therefore, after reducing the dimension of features to overcome the model selection problem and to make treated and non-treated individuals comparable over the period of time the system selected 3447 units (6894 units of observation for two time period) for the analysis. In other words, the algorithm for feature selection selected 6894 units out of 8912 units of (individuals) observation representing QEDs for optimizing the learning equation (\hat{A}). Once, the model estimates the consistent parameter the calculation of ATE and CATE follows the method explained in the model.

Therefore, for empirical analysis the study first satisfy the condition of exogeneity. For this, the study first reduces the dimension of features through feature selection. The study fits sample into sigmoid function (logistic equation), which validates asymptotic properties. By this the study identify the lower and upper bound for the distribution. Thus the proposed study, has adopted a doubles robust approach for estimating treatment effect. The study advocates for a dynamic approach – instead of static approach– to estimate the treatment. This model helps us to estimate the entire distribution for accounting the heterogeneity in treatment effect. For instance, through this model, one can identify how the treatment effect is distributed across individual's group in different regions or geographical areas.

4. Results

As a reference to the above discussed Changes-in-Changes method this section reports the estimated parameter of interest. Following the data-driven methods, essentially satisfying the sufficient assumption of “exogeneity” on database, the training (data) set mimics the QED. In other words blocking all the nodes between treatment variable and potential outcomes makes treatment assignment quasi-random.

Subsequently, estimates of ATE on respective outcomes and CATE on respective variables are based on derivation. The study uses delta method for calculating standard error. In addition, the model has also incorporated the additional covariates⁵⁸ for estimation as it tightens the bounds (Athey and Imbens 2006).

The study proposes the following hypothesis:

Average Treatment Effect

- i. **Null Hypothesis:** There is no difference in probability between the treated and non-treated individuals i.e. $H_0: ATE = 0$
- ii. **Alternative Hypothesis:** $H_1: ATE \neq 0$

Conditional Average Treatment Effect

1) HFSs vs. Non-HFS

- i. **Null Hypothesis:** The average treatment effect is equal in HFSs and Non-HFSs i.e. $H_0: ATE_{HFS} = ATE_{Non-HFS}$
- ii. **Alternative Hypothesis:** The average treatment effect is not equal for HFSs and Non-HFSs i.e. $H_1: ATE_{HFS} \neq ATE_{Non-HFS}$

2) URBAN vs. RURAL

- i. **Null Hypothesis:** The average Treatment effect is equal for rural and urban areas of HFSs and Non-HFSs – both within and between i.e.

$$H_0: ATE_{HFS\#Rural} = ATE_{HFS\#Urban}$$

$$H_0: ATE_{Non-HFS\#Rural} = ATE_{Non-HFS\#Urban}$$

$$H_0: ATET_{HFS\#Rural} = ATET_{Non-HFS\#Rural}$$

$$H_0: ATET_{HFS\#Urban} = ATET_{Non-HFS\#Urban}$$

- ii. **Alternative Hypothesis:** The average treatment effect is not equal for rural and urban regions of HFSs and Non-HFSs – both within and between i.e.

$$H_1: ATE_{HFS\#Rural} \neq ATE_{Non-HFS\#Rural}$$

$$H_1: ATE_{Non-HFS\#Rural} \neq ATE_{Non-HFS\#Urban}$$

$$H_1: ATET_{HFS\#Rural} \neq ATET_{Non-HFS\#Rural} \quad H_1: ATET_{HFS\#Urban} \neq ATET_{Non-HFS\#Urban}$$

Section 2.1 highlights the choice of potential outcome variables that the study focuses for evaluating the intervention in maternal healthcare through NRHM. Thus, the study primarily evaluates the intervention for the following community level expected outcomes identified in policy document.

- a) Institutional Delivery
- b) Ante-Natal Care
- c) Post Natal Care
- d) Safe Delivery

⁵⁸ This includes identification of States, sector (urban or rural), Caste, Assets holding, Education of a women, Number of children, Permission to visit health centre, Group Membership (does the women has any group affiliation with SHGs, Saving/Credit, or Political Party), Exposure to Mass Media, Economic Class (APL or BPL), Acquaintance with doctor or any health worker.



4.1. Institutional Delivery

Table. 1.1 estimates the changes-in-changes (CIC) model explained in section 3. It measure the change (difference) in probabilities of potential outcome i.e. $P_r(ID = 1)$ for treated and non-treated individuals group as time passes from pre-treatment to post-treatment era.

Table.1: Heterogeneous Treatment Effect in Probability Metric Pr (ID=1)⁵⁹

A	Average Treatment Effect	Probability	SE	z value	P>z
	Treated vs Non-Treated	0.5229303** *	0.01930 24	27.0 9	0.0 00
	Conditional Average Treatment Effect (CATE)				
B	HFSs and Non-HFSs				
	Treated#HFS vs Non-Treated#HFS	0.6148785** *	0.02001 49	30.7 2	0.0 00
	Treated #Non-HFS vs Non-Treated#Non-HFS	0.3873512** *	0.24742 2	15.6 6	0.0 00
	Treated#HFS vs Treated#Non-HFS	0.2250082** *	0.01394 98	16.1 3	0.0 00
C	Urban and Rural				
	(Treated#HFS#URBAN) vs (Non-Treated#HFS#URBAN)	0.4692558** *	0.02767 62	16.9 6	0.0 00
	(Treated#HFS#RURAL) vs (Non-Treated#HFS#RURAL)	0.6628019** *	0.02057 01	32.2 2	0.0 00
	(Treated#Non-HFS#URBAN) vs (Non-Treated#Non-HFS#URBAN)	0.2412298** *	0.02593 77	9.3	0.0 00
	(Treated#Non-HFS#RURAL) vs (Non-Treated#Non-HFS#RURAL)	0.4214781** *	0.02656 74	15.8 6	0.0 00
	(Treated#HFS#URBAN) vs (Treated#Non-HFS#URBAN)	0.2301741** *	0.01351 41	17.0 3	0.0 00
	(Treated#HFS#RURAL) vs (Treated#Non-HFS#RURAL)	0.2374404** *	0.01505 35	15.7 7	0.0 00
	(Treated#Non-HFS#RURAL) vs	0.1826078**	0.01556	11.7	0.0

⁵⁹ Significance at ***, **, and *10%; and SE represents the standard error. Also the term # indicates the interaction term.



	(Treated#Non-HFS#URBAN)	*	5	3	00
	(Treated#HFS#RURAL) vs (Treated#HFS#URBAN)	0.1898741**	0.01745	10.8	0.0
		*	11	8	00

Thus, table. 1.1 shows the estimated average treatment effect (ATE) and conditional average treatment effect (CATE). Section. A, of the table asserts that the probability for institutional delivery is higher for individuals in treated group by 52.2 percent than an individuals from untreated group after the policy intervention. Thus, the findings reject the null hypothesis at 1 % significance level that there is no difference in probability between treated and non-treated groups.

Conditional average treatment effect (CATE) for respective states (i.e. between HFSs and Non-HFSs) affirms that treated women were more likely to have institutional delivery than non-treated women in HFSs relatively to Non-HFSs. Furthermore, treated women in HFSs were 22.5 percent more likely to have institutional delivery than treated women in Non-HFSs. In addition, section C in table. 1.1 estimates the CATE for respective geographical location (Urban and Rural) in HFSs and Non-HFSs. First two part of this section affirms that treated women in rural part of the states were more likely to have Institutional delivery than non-treated women in urban areas.

4.2. Ante-Natal Care (ANC)

Ante-natal Care is considered to be the most crucial component for safe motherhood. The probability of maternal and infant mortality increases extensively if a woman has been denied the for ante-natal care check-ups. It is recommended by WHO that at least four ANC check-up is needed for reducing maternal and infant mortality rate.

Table. 2 affirms that the changes in probabilities for at least four ANC check-up between treated and non-treated women over the period of time has not changed much compared to institutional delivery (see table 1.1). However, ATE is equal to 0.062, indicating that women’s in treated group were more likely to have ANC than un-treated women. The study rejects the null hypothesis at 1 % of significance level that there are no differences in changes in probability between treated and un-treated women.

Section B of table. 2 indicate that average treatment effect in High focus states (i.e. ATE_{HFS}) is not equal to average treatment effect in Non-High focus states (i.e. $ATE_{Non-HFS}$). A treated women in HFSs are less likely to have at-least four ANC than a treated women in Non-HFSs by 2.5 percent. Likewise, average treatment effect in HFSs is higher for those women who reside in urban areas than in rural areas, whereas $ATE_{Non-HFS}$ is higher from women in rural areas than in urban location..

Table.2: Heterogeneous Treatment Effect in Probability Metric Pr (ANC=1)



	Average Treatment Effect	Probability	SE	z valu e	P>z
A	Treated vs Non-Treated	0.0621122 ***	0.0234321	2.65	0.008
	Conditional Average Treatment Effect (CATE)				
	HFSs and Non-HFSs				
B	Treated#HFS vs Non-Treated#HFS	0.0651455 ***	0.0233198	2.79	0.005
	Treated #Non-HFS vs Non-Treated#Non-HFS	0.0687034***	0.0271538	2.53	0.011
	Treated#HFS vs Treated#Non-HFS	-0.025512***	0.0050918	-5.01	0.000
	Urban and Rural				
C	(Treated#HFS#URBAN) vs (Non-Treated#HFS#URBAN)	0.0708765 ***	0.026012	2.72	0.006
	(Treated#HFS#RURAL) vs (Non-Treated#HFS#RURAL)	0.0649927 ***	0.0229461	2.83	0.005
	(Treated#Non-HFS#URBAN) vs (Non-Treated#Non-HFS#URBAN)	0.0656568 **	0.0273539	2.4	0.016
	(Treated#Non-HFS#RURAL) vs (Non-Treated#Non-HFS#RURAL)	0.0711522 ***	0.0277534	2.56	0.010
	(Treated#HFS#URBAN) vs (Treated#Non-HFS#URBAN)	-0.0067361	0.0048761	-1.38	0.167
	(Treated#HFS#RURAL) vs (Treated#Non-HFS#RURAL)	-0.0327076***	0.0062892	-5.20	0.000
	(Treated#Non-HFS#RURAL) vs (Treated#Non-HFS#URBAN)	0.0045821 *	0.0026371	1.74	0.082
	(Treated#HFS#RURAL) vs (Treated#HFS#URBAN)	-0.0213893***	0.004792	-4.46	0.000

4.3. Post-Natal Care (PNC)

Post-natal care is a decisive component for attaining the objective of safe motherhood. As per WHO recommendation the women and child should receive post-natal care within 24 hours of delivery. The study has looked at two aspects of post-natal care for evaluating the program; first, did mother and newly born child receive PNC in two months time after delivery and second, was this post natal care was given within 24 hours of delivery.



Table. 3: Heterogeneous Treatment Effect in Probability Metric Pr (PNC2M=1)⁶⁰

	Average Treatment Effect	Probability	SE	z value	P>z
A.	Treated vs Non-Treated	0.2282958***	0.0242501	9.41	0.000
	Conditional Average Treatment Effect				
	HFSs and Non-HFSs				
B.	Treated#HFS vs Non-Treated#HFS	0.2224834***	0.0231254	9.62	0.000
	Treated #Non-HFS vs Non-Treated#Non-HFS	0.2494074***	0.0271339	9.19	0.000
	Treated#HFS vs Treated#Non-HFS	-0.0718259***	0.0093628	-7.67	0.000
	Urban and Rural				
C.	(Treated#HFS#URBAN) vs (Non-Treated#HFS#URBAN)	0.2334764***	0.0244709	9.54	0.000
	(Treated#HFS#RURAL) vs (Non-Treated#HFS#RURAL)	0.2205009***	0.0229202	9.62	0.000
	(Treated#Non-HFS#URBAN) vs (Non-Treated#Non-HFS#URBAN)	0.252525***	0.0280614	9	0.000
	(Treated#Non-HFS#RURAL) vs (Non-Treated#Non-HFS#RURAL)	0.2504019***	0.0269739	9.28	0.000
	(Treated#HFS#URBAN) vs (Treated#Non-HFS#URBAN)	-0.0633384***	0.0093006	-6.81	0.000
	(Treated#HFS#RURAL) vs (Treated#Non-HFS#RURAL)	-0.0759664***	0.0098775	-7.69	0.000
	(Treated#Non-HFS#RURAL) vs (Treated#Non-HFS#URBAN)	-0.0153039**	0.0057774	-2.65	0.008
	(Treated#HFS#RURAL) vs (Treated#HFS#URBAN)	-0.0279319***	0.0103946	-2.69	0.007

⁶⁰ This table represents ATE and CATE for PNC in two months period after the delivery did both – mother and child – has received any post natal care (PNC).



Table. 3 show the ATE and CATE for PNC in two months time after the delivery. The ATE is 0.228 indicating the changes in probability for PNC (for both mother and newly born child) in two months period has increased significantly for treated women relative to un-treated one's. Looking at CATE between HFSs and Non-HFSs, the study uncovers that there is substantial differences in treatment effect between HFSs and Non-HFSs. Accordingly, the study rejects the null hypothesis that there is no differences between ATE_{HFS} and $ATE_{Non-HFS}$. The study reveals that $ATE_{HFS\#Urban}$ is greater than $ATE_{HFS\#Rural}$ in high focus states (See Table. 3), whereas the $ATE_{Non-HFS\#Urban}$ are approximately close to $ATE_{Non-HFS\#Rural}$. In other words, the study affirms that High Focus States has suggestive difference in treatment effect between urban and rural areas than Non-High Focus States.

Now looking at the second aspect of post-natal care i.e. did the mother and newly born child received PNC check up within 24 hours of delivery reveals a bleak picture.

Table. 4: Heterogeneous Treatment Effect in Probability Metric Pr(PNC24=1)

Table 3.4 Changes in Changes Estimator of Treatment Effect in the Probability Metric Pr(PNC24=1)					
A	Average Treatment Effect (ATE)	Probability	SE	z value	P>z
	Treated vs Non-Treated	0.1350729** *	0.028896 8	4.6 7	0.000
B	Conditional Average Treatment Effect (CATE)				
	HFSs and Non-HFSs				
	Treated#HFS vs Non-Treated#HFS	0.1382028** *	0.029144	4.7 4	0.000
	Treated #Non-HFS vs Non-Treated#Non-HFS	0.1302587** *	0.028527 8	4.5 7	0.000
	Treated#HFS vs Treated#Non-HFS	0.0001451	0.000507 2	0.2 9	0.775
C	Urban and Rural				
	(Treated#HFS#URBAN) vs (Non-Treated#HFS#URBAN)	0.1422501** *	0.026845 7	7.7 5	0.000
	(Treated#HFS#RURAL) vs (Non-Treated#HFS#RURAL)	0.1372675** *	0.029185 2	4.7	0.000
	(Treated#Non-HFS#URBAN) vs (Non-Treated#Non-HFS#URBAN)	0.1359452** *	0.029129 2	4.6 7	0.000
	(Treated#Non-HFS#RURAL) vs (Non-Treated#Non-HFS#RURAL)	0.1287383** *	0.028438 6	4.5 3	0.000



(Treated#HFS#URBAN) vs (Treated#Non-HFS#URBAN)	-6.41	0.000081 2	- 0.0 8	0.937
(Treated#HFS#RURAL) vs (Treated#Non-HFS#RURAL)	0.0001935	0.000674 5	0.2 9	0.774
(Treated#Non-HFS#RURAL) vs (Treated#Non-HFS#URBAN)	-0.0001768	0.000615 9	- 0.2 9	0.774
(Treated#HFS#RURAL) vs (Treated#HFS#URBAN)	0.0000231	0.000096 3	0.2 4	0.810

(Fougere & Jaquemet, 2020)

4.4. Safe Delivery⁶¹

Safe delivery incorporates four contributing elements of safe motherhood. As demarcated by WHO ANC, Institutional delivery, PNC and, Immunization are the four contributing elements to obtain safe motherhood for any pregnant women seeking reproductive healthcare services. Therefore, the present study intends to incorporate this idea to estimate the changes in probability of safe delivery between treated and untreated pregnant women since intervention.

Table. 5: Heterogeneous Treatment Effect in the Probability Metric Pr (SAFEDL=1)

A	Average Treatment Effect (ATE)	Probability	SE	z value	P>z
	Treated vs Non-Treated	0.1169***	0.018847 9	6.2	0.000
B	Conditional Average Treatment Effect (CATE)				
	HFSs and Non-HFSs				
	Treated#HFS vs Non-Treated#HFS	0.0906582** *	0.015085 3	6.0 1	0.000
	Treated #Non-HFS vs Non-Treated#Non-HFS	0.1504781 ***	0.024684 7	6.1	0.000
	Treated#HFS vs Treated#Non-HFS	-0.0921466 ***	0.015019 8	- 6.1 4	0.000

⁶¹ This variable is created by using the definition of safe motherhood from WHO. It advocates that there are four contributing element to offer safe motherhood to any pregnant women in need of reproductive healthcare attention. These elements include ANC, Institutional Delivery, PNC and Immunization. Where immunization includes tetanus injection, Iron and Folic acid syrup (or tablets) and vitamin A injection. The study incorporates this idea and constructed a safe delivery variable. Thus the study constructed the variable “Safe Delivery” indication 1 for those women who experienced all the four contributing element of safe motherhood.



C	Urban and Rural				
	(Treated#HFS#URBAN) vs (Non-Treated#HFS#URBAN)	0.1112405** *	0.018815 3	5.9 1	0.000
	(Treated#HFS#RURAL) vs (Non-Treated#HFS#RURAL)	0.0821392** *	0.014066	5.8 4	0.000
	(Treated#Non-HFS#URBAN) vs (Non-Treated#Non-HFS#URBAN)	0.1742957 ***	0.028882 9	6.0 3	0.000
	(Treated#Non-HFS#RURAL) vs (Non-Treated#Non-HFS#RURAL)	0.1448595** *	0.023852 8	6.0 7	0.000
	(Treated#HFS#URBAN) vs (Treated#Non-HFS#URBAN)	- 0.1011709** *	0.016441	- 6.1 5	0.000
	(Treated#HFS#RURAL) vs (Treated#Non-HFS#RURAL)	- 0.0936696** *	0.015233 9	- 6.1 5	0.000
	(Treated#Non-HFS#RURAL) vs (Treated#Non-HFS#URBAN)	- 0.0495067** *	0.012523 6	- 3.9 5	0.000
	(Treated#HFS#RURAL) vs (Treated#HFS#URBAN)	-0.0420055 ***	0.010837 5	- 3.8 8	0.000

Table 1.5 shows the changes in safe delivery between treated and non-treated due to changes in policy over the period of time. It indicates that ATE for safe delivery (due to policy intervention) has been positive 0.1169. Looking at the conditional average treatment effect (CATE) between states indicates that there are momentous differences between HFSs and Non-HFSs. The finding reveals that treated women in Non-HFSs have higher probability for safe delivery than in HFSs (by 0.092). Likewise, the distribution of treatment effect between urban and rural location is significantly diverged. For instance, $ATE_{HFS\#Urban}$ is 0.11 where as $ATE_{HFS\#Rural}$ is 0.082. Likewise, $ATE_{Non-HFS\#Urban}$ is 0.174 and for $ATE_{Non-HFS\#Rural}$ is 0.144. Thus the finding suggests that there is high degree of differences between and within states and raises serious concern for the achieving the goal of sustainable development.

5. Conclusion

The paper tried to bring in recent development in estimating casual effect using Machine Learning tool. This method can estimate complete distribution of counterfactuals and incorporates the scope for estimating heterogeneity in treatment effect, which has been a major challenge for causal inference. The paper has implicitly assumed connectedness for potential outcomes in two-time period. This helps is identifying the Qausi-Experimental Design (QED), representing the natural experiment. In other words, the model use the



intelligence of system to look for units in population database and select that subset of data which satisfies the sufficient condition for casual inference (assumption of exogeneity). The study also suggest that an availability of well defined database with extensive features and continuous time interval in developing countries will help to estimate the impact of key welfare schemes/programs. Data driven methods has several advantages that helps us to learn the existing and design the future policy that can provide better results and utility over the existing ones.

The finding suggests that the program was effective in increasing institutional delivery. It helped in bridging the gap between rural and urban areas between and within HFSs and Non-HFSs. Further investigation predicts that the program was effective to an extent for increasing the utilization of other maternal healthcare outcomes within treated women in aggregate. However, the distribution of treatment effect suggests that treatment effect in Non-HFSs were marginally higher than in HFSs. Moreover, conditional distributions of treatment effect for rural areas in HFSs were relatively lower than in urban areas. Whereas in Non-HFSs the conditional distribution of treatment between rural and urban areas were relatively similar. Although, the finding suggests that the program was effective in increasing the usages of maternal health care in India. However, the conditional distribution of treatment effect between HFSs and Non-HFSs and further, among rural and urban were such that maternal health inequality still prevails.

Bibliography

- Athey, S., & Imbens, G. W. (2006). Identification and Inference in Non-Linear Difference-in-Differences Models. *Journal of Econometrica*, 74(2), 431-497.
- Athey, S., & Imbens, G. W. (2017, Spring). The State of Applied Econometrics: Causality and policy Evaluation. *Journal of Economic Perspectives*, 31(2), 3-32.
- Athey, S., & Imbens, G. W. (2017, Spring). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspective*, 31(2), 3-32.
- Blundell, R., & Dias, M. C. (2009). Alternative Approaches to Evaluation in Empirical Microeconomics. *The Journal of Human Resources*, 44(3), 565-640.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics Methods and Applications*. Cambridge, UK: Cambridge University Press.
- Crabbe, J., Curth, A., & Bica, I. (2022). Benchmarking Heterogeneous Treatment Effect Models through the lens of Interpretability. Retrieved from arXiv:2206.08363
- Debnath, S. (2012, Oct). *Indian School of Business*. (ISB, Producer, & Indian School of Business) Retrieved Jan 2017, from www.isb.edu:
<http://www.isb.edu/sites/default/files/ImprovingMaternalHealth-Doc17120131611.pdf>
- Fougere, D., & Jaquemet, N. (2020). *Policy Evaluation Using Causal Inference Methods*. Bonn: IZA Discussion Papers, No. 12922, institute of Labour Economics (IZA).
- Frankenberg, E. (1995). The effects of access to health care on infant mortality in Indonesia. *Health Transition Review*, 5, 143-63.
- Gol. (2005). *National Rural Health Mission Document*. Government of India, Ministry of Health & Family Welfare. New Delhi: Gol.
- GoMP. (2006). *National Rural Health Mission: Meeting People's Health Needs in Rural Areas*. Government of Madhya Pradesh. GoMP.



- Heckman, J. J., Ichimura, H., & Todd, P. (1998, Oct). Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*, 65, 261-294.
- Hernán, M. A., & Robins, J. M. (2017, March 05). *harvard.edu*. Retrieved April 15, 2017, from <https://www.hsph.harvard.edu/about/>: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2017/03/hernanrobins_v1.10.32.pdf
- Imbens, G. (2003, October). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *NBER Technical Working Paper Series*.
- Jacobs, B., Ir, P., Bigdeli, M., Annear, P. L., & Damme, W. V. (2011, Feb). Addressing access barriers to health services: an analytical framework for selecting appropriate interventions in low-income Asian countries. *Health Policy and Planning*, 1-13.
- Jacobs, L. D., Judd, T. M., & Bhutta, Z. A. (2016). Addressing the Child and Maternal Mortality Crisis in Haiti through a Central Referral Hospital Providing Countrywide Care. *The Permanente Journal*, 20(2), 59-70.
- Jensen, D., Fast, A., Taylor, B. J., & Maier, M. E. (2008). Automatic Identification of Qausi-Experimental Design for Discovering Causal Knowledge. *14th ACM SIGKDD, International Conference on Knowledge and Data Mining*. Las Vegas, NV, USA: 14th ACM SIGKDD.
- Kreif, N., DiazOrdaz, K., Moreno-Serra, R., Mirelman, A., Hidayat, T., & Suhrcke, M. (2022). Estimating Heterogeneous Policy Impacts Using Causal Machine Learning: A Case study of health insurance reforms in Indonesia. *Health Services and Outcomes Research Methodology*, 22, 192-227.
- Kumar, S., & Dansereau, E. (2014). Supply-Side Barriers to Maternity-Care in India: A Facility-Based Analysis. *PLoS ONE*, 9(8).
- McDonagh, M. (1996, March). Is antenatal care effective in reducing maternal morbidity and mortality? *Health Policy and Planning*, 11(1), 1-15.
- Ng, A. (2008, July 22). *stanford.edu*. Retrieved Jan 10, 2017, from <http://cs229.stanford.edu:> <http://cs229.stanford.edu/notes/cs229-notes5.pdf>
- Ng, A. (2008, July 22). *Stanford.edu*. Retrieved Jan 7, 2017, from <http://cs229.stanford.edu/materials.html>: <http://cs229.stanford.edu/notes/cs229-notes1.pdf>
- Ng, A. (2008, Jul 22). *Stanford.edu*. Retrieved Jan 10, 2017, from <http://cs229.stanford.edu:> <http://cs229.stanford.edu/notes/cs229-notes4.pdf>
- Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., & Gama, J. (2022, March). Methods and Tools for causal Discovery and causal Inference. *WIREs Data Mining and Knowledge Discovery*, 1-39.
- Noordam, A. C., Kuepper, B. M., Stekelenburg, J., & Milen, A. (2011, May). Improvement of maternal health services through the use of mobile phones. *Tropical Medicine & International Health*, 16(5), 622-26.
- Olusegun, L., Thomas, R., & Micheal, I. (2012). Curbing maternal and child mortality: The Nigerian Experience. *International Journal of Nursing and Midwifery*, 4(3), 33-39.
- Oyerinde, K. (2013). Can Antenatal Care Result in Significant Maternal Mortality Reduction in Developing Countries. *Comunity Medicine & Health Education*, 3(2).



- Panel, A. P. (2010, September). Maternal Health: Investing in the Lifeline of Healthy Societies & Economies. Retrieved December 2016, from who.int:
http://www.who.int/pmnch/topics/maternal/app_maternal_health_english.pdf
- Pearl, J. (2009). Causal Inference in Statistics: An Overview. *Statistics Survey*, 3, 96-146.
- Rai, K. R., & Singh, K. P. (2012, Oct-Dec). Janani Suraksha Yojana: the conditional cash transfer scheme to reduce maternal mortality in India- A need for reassessment. *WHO South-East Asia Journal of Public Health*, 1(4).
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387), 516-524.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387), 516-524.
- Shi, J., & Norgeot, B. (2022, July). Learning Causal Effects From Observational Data in Healthcare: A review and Summary. *Frontiers in Medicine*.
doi:doi.org/10.3389/fmed.2022.864882
- Sines, E., Syed, U., Wall, S., & Worley, H. (2007, January). Postnatal Care: A Critical Opportunity to Save Mothers and Newborns. POPULATION REFERENCE BUREAU.
- Singh, A. (2016, Aug). Supply-side barriers to maternal health care utilization at health sub-centers in India. *Peer J*, 3(4).
- Smith, J. (2022). Treatment Effect heterogeneity. IZA Discussion Paper Series.
- UNICEF. (2009). *The State of World's Children*. United Nations. New York: United Nations Children's Fund.
- Walraven, G., Telfer, M., Rowley, J., & Ronsmans, C. (2000). Maternal mortality in rural Gambia: levels, causes and contributing factors. *Bulletin of the World Health Organization*, 78(5).
- WHO. (2013). *WHO recommendations on Postnatal care of the mother and newborn*. World Health Organization. WHO.
- WHO. (2016). *WHO recommendations on antenatal care for a positive pregnancy experience*. World Health Organization. WHO.

